International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(9s) 2026 pp. 188-200
https://doi.org/10.70917/ijcisim-2026-3501
Article

# Enhanced Facial Expression Recognition Using Multi-Scale Attention Fusion and Deep Learning

**Ali Naji Salami[1, *], Abdul Jabbar P[2], Anuj Mohamed [3]**

[1,3] School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
[2] Graduate School, Stamford International University, Prawet,10250, Bangkok, Thailand
[1]asdbabel87@gmail.com,[2]abduljabbar.perumbalath@stamford.edu ,[3]anujmohamed@mgu.ac.in

**Abstract:** Facial Expression Recognition (FER) is an important aspect of human-computer interaction, where machines can read through the facial expression to understand how an individual feels. Although the deep learning field has made great progress, it is difficult to detect delicate and contextual emotions in real life since the lighting, pose, occlusions, and low-resolution images are not constant. To overcome these limitations, this paper suggests a Multi-Scale Attention Fusion Network (MSAFN) with an EfficientNet-B1 backbone. This is because we are using three parallel convolutional branches with different kernel sizes of 1x1, 3x3, and 5x5 to capture fine-grained local information, mid-level structural information, and broader contextual information. The attention-based fusion of the extracted multi-scale features is followed by the classification of the features by fully connected layers with the help of dropout regularization. We tested our model on three additional well-known standard benchmark datasets: FER2013, RAF-DB, and AffectNet. We equalized emotion types to the seven universal emotions that Ekman proposed: anger, disgust, fear, happiness, sadness, surprise, and neutral. With accuracies of 72.18% on FER2013, 86.96% on RAF-DB, and 78.93% on AffectNet, the MSAFN outperformed cutting-edge models like ConvNeXt and EfficientNet baselines. Our findings show that multi-scale attention can be used to capture global facial geometry and local action units, which can be used to make them robust in varying visual conditions. The proposed architecture shows a significant improvement in detecting fine-grained emotions in challenging real-world situations with high intra-class variation and high noise.

**Keywords:** Facial Expression Recognition, Multi-Scale Attention Fusion, EfficientNet-B1, Deep Learning, CNN, Emotion Classification, Computer Vision, Human-Computer Interaction

## 1. Introduction

Facial expression is one of the most important forms of nonverbal human communication, providing important clues to what people are thinking, feeling, and intending. In recent years, computer vision and Artificial Intelligence have evolved, and “FER (Facial Expression Recognition)” is an emerging key area of research, enabling computers to interpret facial expressions automatically and improve the response and naturalness of Human Computer Interaction (HCI) [1]. The FER technologies are very useful in various fields of application, such as in health care and psychological assessment [2], in education, for tracking a student's engagement level in the learning process [3] and in smart transportation where the emotional state of a driver can be determined to ensure safer driving experiences [4].

in the 1970s, Ekman and Friesen first described 6 basic emotions (happiness, anger, surprise, fear, disgust, sadness). Later, Contempt was identified as another basic emotion and hence seven emotion categories have formed the basis of many later emotion recognition studies [5]. There has been a significant methodological advancement in FER. Most of the previous methods used handcrafted feature descriptors, such as Local Binary Patterns (LBP) [1] and Histogram of Oriented Gradients (HOG) [2] as well as widely used machine-learning classifiers, especially Support Vector Machines (SVM) [6]. Although these methods provided a starting point, they were always constrained by the use of real-world facial data, and were sensitive to changes in occlusion, head pose and illumination. The most widely used FER design is the convolutional neural network (CNN), which has proved to be a game-changer in the field by automatically extracting features from raw pixels and stimuli, enhancing performance [7].

Despite these advances, there are huge challenges that persist, especially in the case of in-the-wild expression recognition. The practical imagery and video add hundreds of complications, such as high intra-class variability, i.e., varying subjects expressing the same emotions differently, and minor inter-class similarities, i.e.,

the subtle distinctions between fear and surprise. To minimize such complications, scientists have increasingly tried to integrate attention systems, which are based on the selective attention of the human eye to interesting regions of the scene [8]. One of the most interesting findings of the new literature is the understanding of the fact that the phenomenon of facial expressions is inherently multi-scale. Global characteristics define large-scale affective contexts, while subtle affective states are defined by small-scale, regional characteristics, such as eyes, eyebrows, and mouth [9]. In turn, there has now been a new avenue that can be promising, namely, the combination of multi-scale feature extraction and attention. This method is often known as Multi-Scale Attention Fusion (MSAF), which is designed to learn and fuse facial features between different receptive fields and granularities, therefore, building a more robust and resilient representation [10].

The "multi-scale attention fusion network (MSAFN)" in this study uses EfficientNet-B1 to extract and fuse facial features across scales. The model was trained and tested using three common facial emotion recognition datasets: FER-2013, AffectNet, and RAF-DB.There are four primary contributions to this work. In order to extract fine textures, mid-level structural features, and more comprehensive contextual information at the same time, we first present a multiscale architecture with three parallel convolutional branches (1×1, 3×3, and 5×5). Secondly, we incorporate a multiscale convolutional fusion module to improve the aggregated features' ability for discrimination. Third, we test the suggested model on three popular datasets, showing competitive performance and good generalizability in contrast to the most recent methods. Finally, we present a comprehensive empirical analysis and discussion, highlighting the effectiveness of the multi-scale attention strategy and its contributions to enhancing FER performance under real-world conditions.

## 2. Related work

This section discusses the facial expression recognition datasets to be used in our experiments and analyzes hand-crafted features and deep-learning-based methods in past studies. We pinpoint the deficiencies of the current methods, which form the basis of our multi-scale-based attention approach. The performance of AI models is mostly dependent on public datasets. The earlier researches have shown that a majority of the datasets obtained in the laboratory setting produce good training performance due to the homogeneity and repetitive nature of the samples, but cannot effectively classify emotions in new data and on real-world images [11]. Also, class imbalance, i.e., a difference between the number of samples in one class and another, is one of the most severe issues in databases, as it biases certain models to favor the more representative one. We have chosen AffectNet [12], RAF-DB [13], and FER-2013[14] as datasets to be used in our study.

FER is a popular dataset with 35,887 images (28,709 training, 3,589 validations, 3,589 testing). Anger, disgust, fear, happiness, sadness, surprise, and neutral comprise FER2013. These categories are then ranked from 0 to 6. Zeng et al. [15] achieved 61.86% accuracy by combining hand-crafted features with deep learning on the FER-2013 dataset. Sarvakar et al. [16] obtained about 54-55 % of accuracy by using a CNN with the FER-2013 dataset. Mozaffari et al. [17] applied FER-2013 to VGG16 and EfficientNet-B1 without any additional training, which showed a test accuracy of 72 %. AffectNet is a massive dataset of faces that is an aggregation of internet-collected images. Anger, disgust, fear, happiness, neutral, sadness, and surprise are among AffectNet-7's 287,401 manually marked and annotated images. Training and validation data were separated. The article by Gomez-Sirvent et al. [18] was a study that used AffectNet to test their Voting Residual Network, using a modified ResNet-18 architecture that demonstrated an accuracy of 63.06% on AffectNet with seven emotion categories.

In a different work, Yen and Li [19] used the AffectNet dataset and tested transfer learning CNN models of ResNet-50, Xception, EfficientNet-B0, Inception, and DenseNet-121 in general, with an overall accuracy of around 61%. The RAF-DB is a facial expression dataset that includes 29,672 online images that were evaluated by 40 participants of various ages, genders, and ethnicities. It attempts to assess both simple and complex emotions through real-life facial expressions [20]. Qutub and Atay used the RAF-DB dataset to train with deep learning CNN architectures, such as ResNet18, VGGNet16, VGGNet19, and EfficientNet-B0. Of them, the most successful approach is ResNet18 in the test accuracy of 86.02%, as it shows the strength of deep CNNs in real-world facial emotion recognition [21]. Moreover, Gursesli et al. relied on the RAF-DB database and the customized Lightweight CNN Model (CLCM) that is founded on the MobileNetV2 architecture, obtaining an accuracy of 84 % of facial emotion recognition [22]. The initial FER experiments used handcrafted feature descriptors to learn the delicate changes in facial muscle variations in a two-step process: extracting the discriminative features and presenting them to classical machine learning classifiers.

Other methods for encoding texture and edges include the HOG and Local Binary Patterns (LBP), which are typically used when combined with SVM [6], [23]. However, due to pose, light, and obstruction problems, these approaches were characterized by poor generalization in the real world. Deep learning, and more especially CNNs, revolutionized FER by enabling end-to-end learning and automatic extraction of hierarchical features [24]. Other architectures like AlexNet[25], VGG [26], GoogLeNet[27], DenseNet[28], and Inception networks improved performance significantly, and deeper and more accurate models became possible with the residual

connections of ResNet[29]. On large and diverse data sets like FER-2013, AffectNet, and RAF-DB, deep learning-based systems perform better. Mobile device and internet use have increased, leading to more lightweight models.Although the deep CNNs turned out to be effective feature extractors, researchers discovered that they had one major limitation, where all spatial locations and feature channels are treated equally, unlike how human faces are perceived, whereby we tend to pay attention to important objects such as eyes, mouth, and eyebrows. It resulted in adding attention mechanisms to FER models, which enables networks to dynamically make weightings of the importance of various input components.

There are a few types of attention that can be considered in FER. Spatial attention gives more emphasis to the most informative parts of the face, producing attention maps which focus on the important landmarks and inhibit the irrelevant background, enhancing the strength of occlusions and changes in poses [30]. Channel attention identifies the channel weights to re-emphasize what to focus on, boosting recommendations of discriminative features like smiles or frowns [31]. The model suggested by Ghadai et al. is called FCMSA-AF (Feature Complementation and Multi-Scale Attention with Attention Fusion) and consists of a two-branch multi-scale attention network. One of the innovations is the Feature Complementation Module (FCM) that models the dependency between the two branches to make sure that they are learning complementary features (instead of redundant features). The last features are then merged with another attention layer [32]. The SFER-MDFAE model was specifically constructed by Shou et al. [3] to recognize the expression of students in the smart classroom setting. The key region-oriented attention mechanism emphasizes detailed information, while the multi-scale dual-pooling feature aggregation module extracts data across various scales.

## 3. Methodology

As shown in Figure 1, we used three models in this study: ConvNeXt, EfficientNet-B1, and our suggested MSAFN. These models have been trained and evaluated on FER-2013, RAF-DB, and AffectNet-7. Each emotional category's accuracy, precision, recall, as well as F1-score were assessed. Accuracy, precision, recall, and F1-score are the harmonic means of correctly classified samples, identified positives, and actual positives. In order to improve the robustness and generalizability, all models were trained with original data splits on each dataset, in order to compare with previous studies. The following sections contain detailed performance analysis for each of the models on all the datasets, followed by comparative performance with state-of-the-art methods.

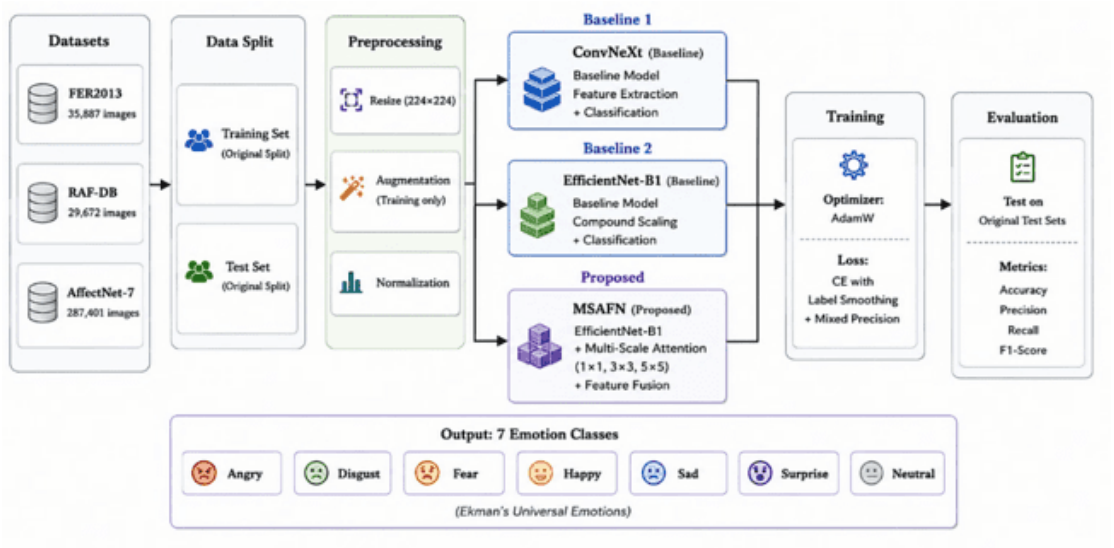


Figure 1. The Proposed Architecture

### *3.1 Model Architecture*

This section describes this study's architectures. EfficientNet-B1 and ConvNeXt-Base were used, as well as the MSAFN model, built on EfficientNet-B1 as its backbone. Figure 1, is related to our methodology in which three datasets of benchmarks (FER2013, RAF-DB, AffectNet-7) are processed and fed into ConvNeXt, EfficientNet-B1 base lines, as well as our proposed multi-scale attention fusion base MSAFN (1x1, 3x3, 5x5

kernel). Correctly classified samples, identified positives, and actual positives are harmonic means of accuracy, precision, recall, and F1-score.

#### 3.1.1 EfficientNet-B1

EfficientNet was unveiled by Google AI in 2019. It offers an exciting step in the domain of neural networks. These convolutional models aim for a smart mix: great accuracy and efficient computing, actually. The main new concept is the compound scaling method. This method elegantly changes the important sizes of the network. Network width means channels within a layer. It impacts how well the model takes in diverse aspects of its task. Depth is the number of layers. It indicates how fully the model learns advanced, or hierarchical ideas. The resolution is about the input photos. This determines image detail and the detail that's available for extracting qualities. Instead of isolated adjustments, EfficientNet smoothly adjusts them for a balanced output that is precise but economic. The underlying parts are based on MBConv elements. In addition, the Squeeze-and-Excitation elements are included. They give an edge to quality evaluation and reduce needless operations.

EfficientNet now shows record levels of task completion. Plus, that outcome occurs at limited costs, in terms of computation. Moreover, you can choose between variants like B0, B1, up to B7. They allow you to pick the most efficient version. Picked regarding computer capabilities or regarding the challenge presented. Because of that, EfficientNet is popular. It finds its niche where balance is required; balance of efficiency with strong model output for computer vision purposes. Using a fixed set of coefficients, the EfficientNet-B1 technique uniformly changes all network dimensions, including depth, width, and resolution, using a simple and remarkably effective scaling factor. Compared to other CNN models, EfficientNetB1 typically achieves greater accuracy with fewer parameters.

#### 3.1.2 ConvNeXt-Base

Contemporary Convolutional Neural Networks. A modernized convolutional neural network, ConvNeXt [32], introduced by Liu et al. in 2022, is made to compete with transformer-based models, especially Vision Transformers (ViT), on computer vision tasks. ConvNeXt-Base has a hierarchical architecture, comparable to ResNet, but with several improvements. The kernel size is 7x7 to capture long-range dependencies, depthwise convolution blocks reduce computational complexity without affecting representational capacity, layer normalization replaces batch normalization to improve training stability, activation and convolution are decoupled, and hierarchical feature scaling with successively smaller spatial resolution as well as deeper channel multiplication in four network stages is used. By doing so, ConvNeXt-Base has demonstrated competitive or better image performance than Swin Transformer on ImageNet- 1K, outperforming previous CNN models such as ResNet-50 and EfficientNet, as well as being trained and inferred more efficiently than transformer-based models.

#### 3.1.3 Multi-Scale Attention Fusion (MSAFN) with EfficientNet-B

To improve the model's capacity to learn emotional cues across various scales, we employ in our study the Multi-Scale Attention Fusion (MSAFN) module, which is integrated after the backbone feature extraction stage. The backbone network we use is the EfficientNet-B1 architecture, which shows a good capability of representation and efficiency, serving as a good basis for the extraction of high-level semantic facial features. So, we begin by using an EfficientNet-B1 model that has been pre-trained on the ImageNet dataset, which is a powerful feature extractor that extracts and processes the input face image and produces rich feature maps that represent high-level concepts such as edges, textures, shapes, and facial components (including nose, mouth, eyes).

The multi-scale processing of the features is the core innovation of the MSAFN module, since instead of using just the final feature map, we take an intermediate feature representation and process it through three parallel convolutional streams with kernel values of 1×1, 3×3, 5×5. The 1×1 convolution is designed to detect anything related to channel-specific information, and the convolutions 3×3 convolutions are designed to extract local contextual features that lead to the recognition of textures related to expressions. The 5×5 convolutions serve the purpose of processing the global structural shots of the facial regions, which involves thinking about the coordination between the openness of the eyes and the curvature of the mouth, when distinguishing different similar expressions such as fear and surprise. See figure 2.

Through processing features simultaneously at three different scales, the model is able to attend to various aspects of the face, and the feature maps from these parallel convolutional layers are concatenated to yield a single unified feature tensor containing information from all three scales. This fused multi-scale representation is fed out to the classification head to produce emotion prediction output using adaptive average pooling, which compacts the intermediate feature representation spatial dimensions. This vector passes through fully connected layers with ReLU activation and Dropout regularization to tune the neural network functional mappings to

discriminate the classes of interest and prevent overfitting, then feeds into the final fully connected layer to produce class logits corresponding to the target emotion labels, which are softmax normalized to produce the emotion probability distribution.

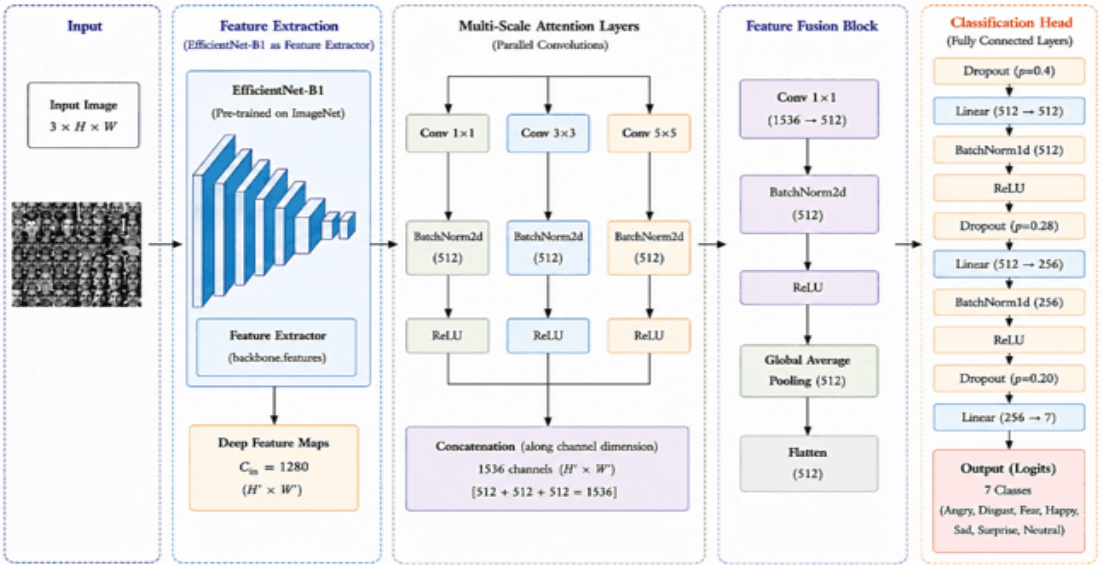


Figure 2: Multi-Scale Attention Fusion (MSAFN) Architecture

## *3.2 Datasets*

This study tested our proposed model on three widely researched facial expression recognition (FER) databases to determine its effectiveness, generalizability, and robustness across various visual and contextual variables: FER-2013, RAF-DB, and AffectNet. The FER-2013 database is an automatically-created data compilation from impressions from Google image searches, and presents difficulties, namely, grayscale limitation, annotation noise, and appearance variations such as positions of the faces and occlusion. The RAF-DB stands for Real-world Affective Face Database and was created by the group effort of analysis by different human annotators, along with a "majority vote" practice. It, like the others, provides significantly real-world variations in light, head position, and intensity of expression. Nevertheless, the RAF-DB suffers greatly from the practice of annotation imbalance, especially as this follows through to the "fear" classification. AffectNet (one of the largest FER databases) comprises more than 400,000 manually labelled images from the internet, and presents some of the same problems, such as the age effect involved in image quality differentials, variation in resolutions, and in posing variation and overlap in expressions.

Figure 3 shows a sample from each dataset used, with one image representing each category. Although the original AffectNet has eight affective categories, such as contempt, and RAF-DB has 11 more categories of emotions, the current study balances the affective taxonomy of all data sets with the 7 universal emotions of Ekman, who organizes them into anger, disgust, fear, happiness, sadness, surprise, as well as neutral. This unification ensures a consistent evaluation framework, reduces labelling inconsistencies, and facilitates reliable cross-database comparison of model performance.


Figure 3: Facial expression images of different datasets (A) FER-2013, (B) RAF-DB and (C) AffectNet

## 4. Model Training and Testing

The proposed training and evaluation pipeline is shown in Figure 1, which consists of data preprocessing and augmentation, model training, and model evaluation.

### *4.1 Preprocessing and data augmentation*

Data preprocessing is necessary for the effective facial expression recognition deep learning models. We created a pipeline for image processing to optimize the data preparation for our neural network model. This procedure takes several crucial steps to ensure data uniformity, diversity, and model learning and generalization. The first step is to resize all images to a fixed size of 224×224, since deep learning models require images of the same dimensions. After this, we randomly applied a set of data augmentation techniques to the training set, to artificially increase the diversity of the data and prevent overfitting. In particular, we used random horizontal flipping with a probability of 0.5 to simulate the left-right variations in the facial orientation, random rotation up to 15 degrees to account for minor pose changes, and color jittering to adjust the brightness, contrast, saturation, and hue to account for illumination inconsistencies.

Furthermore, we used random affine transformations (small translations, scaling variations) to make the system tolerant to spatial shifts and face size variations, random perspective distortion to simulate depth-related changes and camera angle variations, and random grayscale conversion to minimize color dependency. ImageNet mean and standard deviation values have been employed to normalize and convert images to tensor format after augmentation. Lastly, we applied random erasing to simulate partial occlusions and trained the model to recognize the facial features even if they are partially covered. For the purpose of assessing the performance of the model on unseen data that represents real-world conditions, we simplified the processing of the validation dataset to dimension standardization, conversion to tensors, and normalization without data augmentation.

### *4.2 Training and Optimization Procedures*

The number of images used varies between the different datasets. The training and testing sets were not changed from the original number of images in the AffectNet-7, RAF-DB and FER-2013 databases, as recommended by the creators, to allow comparison with previous studies. We trained all models separately on each dataset and tested them on the corresponding testing sets. Our proposed system employs an enhanced facial emotion recognition architecture built upon the EfficientNet-B1 backbone, selected for its compound scaling strategy that balances network depth, width, and input resolution efficiently. We replaced the conventional classification layers with a Multi-Scale Attention Fusion Network (MSA-EfficientNet) to enhance recognition of subtle emotional features. Following feature extraction, high-level representations pass through three parallel attention branches utilizing 1×1, 3×3, and 5×5 convolutional kernels, capturing feature dependencies at different receptive field scales.

After 1x1 convolution, multi-scale feature maps are batch normalized and ReLU activated. Adaptive Average Pooling compresses a classification head's fully connected layers' features with Batch Normalization, ReLU activation, and Dropout as shown in Figure 2. The AdamW optimizer with Cross-Entropy Loss and label smoothing to prevent overconfidence was used to train, while optional Focal Loss highlighted misclassified examples for class imbalance. The ReduceLROnPlateau scheduler dynamically adapts the learning rate based on validation accuracy. To save memory, we used mixed precision training with autocast() and GradScaler(), and early stopping to stop training when the loss stops improving.

### *4.3 Evaluation Procedures*

We assess the classification accuracy of the model and the confusion matrix, which displays the model's predictions for each emotion category and identifies its strengths and misclassification patterns. Precision, recall and F1 score are computed per class. The model's accuracy is its overall correctness, while precision is the ratio of true positive predictions to all positive predictions. Recall is the percentage of positive events that are correctly identified. High metrics mean good model performance. The harmonic mean of precision and recall has been determined by the F1 score. Like a confusion matrix, large numbers on the diagonal indicate correct classifications, and large numbers off the diagonal indicate incorrect classifications between the classes. This comprehensive evaluation framework guarantees a thorough and accurate evaluation of model's performance in real-world facial emotion recognition scenarios.

## 5. Result and Discussion

The performance of the three deep learning models evaluated in this section is analysed on three benchmark facial expression recognition datasets: AffectNet, FER2013 and RAF-DB. The analysis is conducted using precision, recall, and F1-score metrics, providing a detailed breakdown of the accuracy of each emotion

class, and using overall classification accuracy, a fine-grained comparison of the performance of each architecture on each emotion class is made.

### *5.1. Performance Evaluation on AffectNet Dataset*

The AffectNet dataset is faced with significant difficulties because of variability of images in the real world and the delicacy of the exhibited emotions. As demonstrated by the performance metrics that are compiled in Table 1), variations were discovered in both the overall accuracy and the capacity to classify distinct affective categories in each of the three models evaluated.The ConvNeXt architecture achieved an overall accuracy of 77.58%, demonstrating strong performance in recognizing emotions characterized by clear facial cues, including happiness (F1 = 91.65%) and surprise (F1 = 77.31%). However, according to the confusion matrix shown in figure 5, the model showed low performance when it comes to low-intensity states, including those in the neutral category (Recall = 67.79%). This weakness is also captured in the disproportionate precision-recall characteristic of fear class, i.e., high precision (84.83%) and low recall (71.67%), meaning that fear is accurately identified only when the indicators are conspicuously salient. Figure 4 attributes of architectural performance have further revealed a noticeable discrepancy between training and validation curves of ConvNeXt as compared to its architectural counterparts.

Table 1: Performance metrics of the models on the AffectNet dataset.

| **AffectNet** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Covnxet | | | Effecientnet | | | MSAFN | | |
| Emotion | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| angry | 77.43% | 76.42 % | 76.92 % | 77.92% | 78.60 % | 78.26 % | 77.97% | 79.33 % | 78.64 % |
| disgust | 74.18% | 73.26 % | 73.72 % | 78.63% | 69.40 % | 73.73 % | 74.43% | 75.58 % | 75.00 % |
| fear | 84.83% | 71.67 % | 77.70 % | 82.92% | 73.69 % | 78.04 % | 83.49% | 74.20 % | 78.57 % |
| happy | 90.12% | 93.23 % | 91.65 % | 89.96% | 91.58 % | 90.76 % | 87.98% | 93.53 % | 90.67 % |
| neutral | 72.20% | 67.79 % | 69.92 % | 65.15% | 72.42 % | 68.59 % | 72.23% | 70.11 % | 71.15 % |
| sad | 71.01% | 74.06 % | 72.50 % | 76.58% | 71.38 % | 73.89 % | 76.95% | 74.06 % | 75.48 % |
| surprise | 73.47% | 81.58 % | 77.31 % | 72.70% | 82.30 % | 77.20 % | 77.73% | 81.21 % | 79.43 % |
| accuracy | 77.58% | | | 77.67% | | | 78.93% | | |

On the other hand, EfficientNet also showed a slight change in total accuracy (77.67%), reaching higher levels of identification of any nuanced muscular activity in the sad (F1 = 73.89%) and fear (F1 = 78.04%) groups. However, the neutral expression was one of the bottlenecks of the performance ( F1 = 68.59%), and the challenge of separating subdued facial states clearly persisted. Figure 4 demonstrates the stability of the model by the learning curves provided, which show a more aligned convergence trend.

This was improved over the baseline architectures with the MSAF-Fusion model hitting up to 78.93% accuracy and across all difficult categories, that include anger (F1=78.64%), fear (F1=78.57%), and surprise (F1=79.43%). The model also performs well in fine-grained spatial cues that are necessary in distinguishing complex expressions by using multi-scale attention to combine both global structure and localized facial areas. As can be seen by the diagonal dominance of the confusion matrices in figure 4 and convergence curves in figure 4, the MSAFN architecture is highly robust and stable to learning, which successfully addresses overlaps of emotional categories.

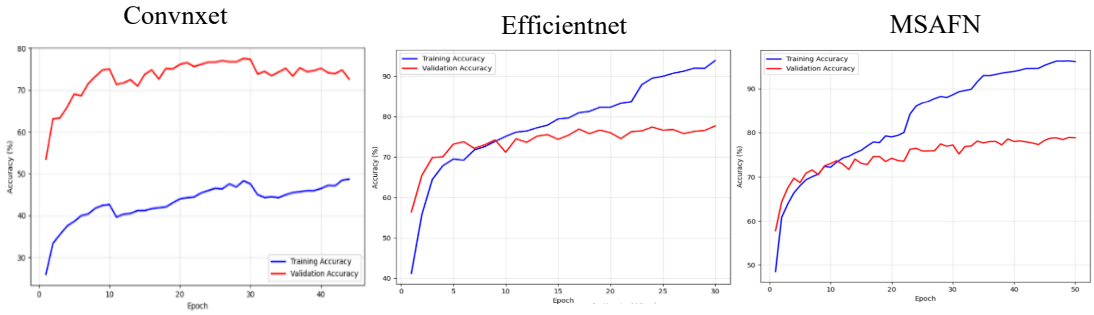


Figure 4: Training and validation accuracy curves for the proposed models on AffectNet dataset

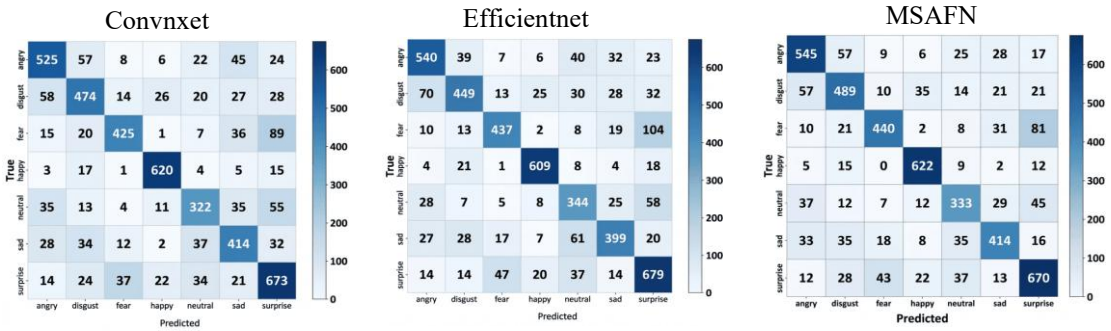


Figure 5: Confusion matrix for the proposed models on AffectNet dataset

## *5.2 Performance Evaluation on FER-2013 Dataset*

The low resolution and grayscale nature of the FER-2013 dataset makes emotional cues less robust and more ambiguous visually, reducing the discrimination task performance of all models, as shown in Table 2, which compares architecture performance on FER2013 data.

Table 2: Performance metrics of the models on the FER-2013 dataset.

| **FER2013** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Covnxet | | | eFFecientnet | | | MSAFN | | |
| Emotion | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| angry | 60.88% | 67.75% | 64.13% | 61.56% | 64.20% | 62.85% | 63.29% | 69.10% | 66.07% |
| disgust | 77.46% | 49.55% | 60.44% | 69.05% | 78.38% | 73.42% | 85.26% | 72.97% | 78.64% |
| fear | 63.36% | 44.92% | 52.57% | 60.30% | 50.88% | 55.19% | 60.84% | 55.08% | 57.82% |
| happy | 89.89% | 89.68% | 89.79% | 87.58% | 89.80% | 88.67% | 88.88% | 89.63% | 89.25% |
| neutral | 62.90% | 75.91% | 68.80% | 65.50% | 70.07% | 67.71% | 65.91% | 70.40% | 68.08% |
| sad | 60.17% | 58.38% | 59.26% | 60.26% | 58.62% | 59.43% | 62.47% | 58.86% | 60.61% |
| surprise | 80.52% | 82.07% | 81.29% | 80.67% | 81.35% | 81.01% | 82.27% | 82.07% | 82.17% |
| accuracy | 71.06% | | | 70.87% | | | 72.18% | | |

ConvNeXt model had a total accuracy of 71.06%. Although it presented strong results in high-variance items like happy (F1 = 89.79%) and surprise (F1 = 81.29%), table 2 indicates that there was a significant decrease in recalling fear (44.92%) and disgust (49.55%). This shows that the traditional modeling in ConvNeXt is not effective when dealing with the fine, low-resolution expression patterns of this dataset.

The total accuracy of the EfficientNet model was 70.87%. Even though its general accuracy was a little worse than ConvNeXt, its feature representation of certain classes was more efficient, which raised the disgust F1-score to 0.7342. Nevertheless, the confusion matrix in figure 7 indicates that such categories as fear (F1 = 55.19%) and sad (F1 = 59.43%) are still difficult to learn because there are not enough texture details of grayscale

images.

Multi-Scale Attention Fusion Network (MSAFN) proposed was found to be superior to both baseline models and recorded the highest overall accuracy of 72.18%. The high performance of MSAFN is specially experienced in its capacity to categorize the complex emotions; it got a lot higher F1-scores in disgust (76.40%) and angry (66.07%) in comparison with its counterparts. As indicated in the confusion matrix (Figure 7) and the stability of its accuracy curves (Figure 6), finer-grained emotional cues are better represented by the integration of multi-scale attention mechanisms by the model. These results indicate that MSAFN could be successfully used to counteract the deficiency of visual clarity in FER2013 and offers discriminative strength than typical ConvNeXt or EfficientNet designs.

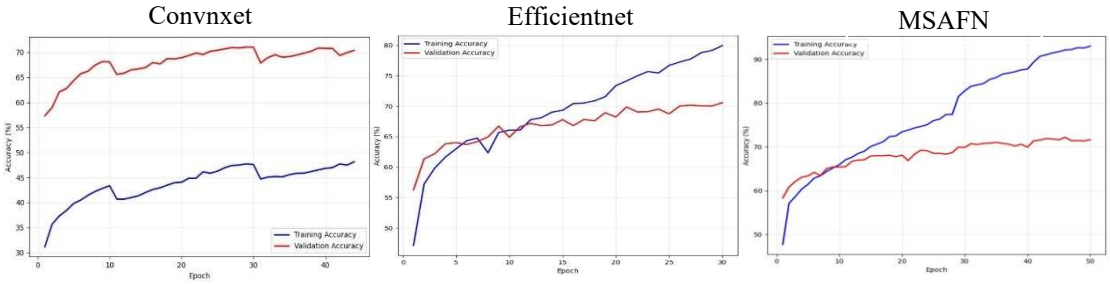


Figure 6: Training and validation accuracy curves for the proposed models on FER-2013 dataset

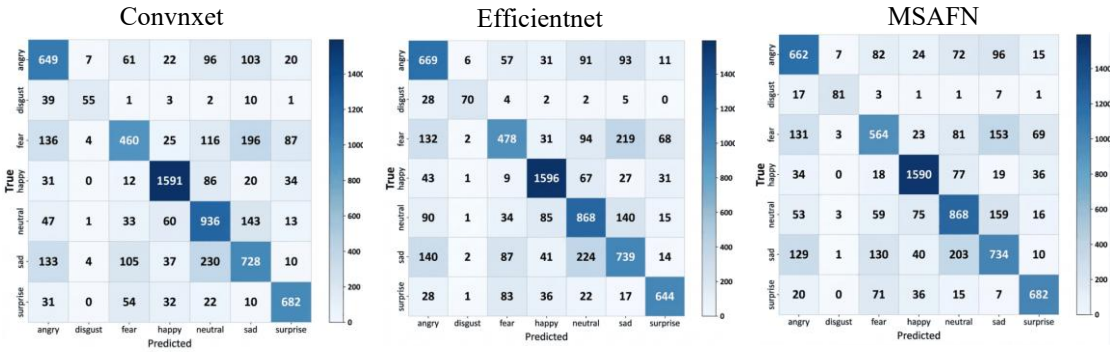


Figure 7: Confusion matrix for the proposed models on FER-2013 dataset

## *5.3 Performance Evaluation on RAF-DB Dataset*

Better-quality images and more obvious emotional indicators are available in the RAF-DB dataset. The findings show that the three architectures' classification accuracy is gradually improving. Table 3 shows evaluation performance on RAF-DB.

| **RAF-DB** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Covnxet | | | eFFecientnet | | | MSAFN | | |
| Emotion | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| angry | 72.58% | 83.33% | 77.59% | 78.31% | 80.25% | 79.27% | 74.59% | 85.19% | 79.54% |
| disgust | 56.80% | 60.00% | 58.36% | 68.75% | 55.00% | 61.11% | 67.15% | 57.50% | 61.95% |
| fear | 70.00% | 56.76% | 62.69% | 67.61% | 64.86% | 66.21% | 80.00% | 48.65% | 60.50% |
| happy | 96.17% | 91.14% | 93.59% | 97.08% | 92.49% | 94.73% | 95.82% | 92.74% | 94.25% |
| neutral | 80.32% | 87.65% | 83.83% | 84.25% | 83.38% | 83.81% | 82.81% | 88.53% | 85.57% |
| sad | 82.04% | 84.10% | 83.06% | 78.92% | 91.63% | 84.80% | 84.98% | 86.40% | 85.68% |
| surprise | 89.93% | 81.46% | 85.49% | 83.53% | 87.84% | 85.63% | 84.46% | 87.54% | 85.97% |
| accuracy | 85.37% | | | 86.57% | | | 86.96% | | |

Table 3: Performance metrics of the models on the RAF-DB dataset.

The ConvNeXt model demonstrated the accuracy of 85.37 doing exceptionally well in the high-clarity classes, like happy (F1 = 93.59%), neutral (F1 = 83.83%), and sad (F1 = 83.06%). Nevertheless, the fear category followed by 56.76% recall, even though it seemed that there was a persistent ambiguity when emotional states had similar muscle actions on the face.The accuracy of the EfficientNet model was greater, at 86.57%. This model achieved the balance between the precision and recall in most classes. It should be noted that it was found that performance improvement occurred in both sad (F1 = 84.80%) and surprise (F1 = 85.63%) categories and this indicates the high ability of the model to generalize when the visual features are different.Multi-Scale Attention Fusion Network (MSAFN) achieved the highest accuracy of 86.96% (Table 2).

The advanced discriminatory ability of MSAFN is also noted as it is best in identifying neutral (F1 = 85.57%), sad (F1 = 85.68%), and surprise (F1 = 85.97%) expressions. As indicated in the confusion matrix (Figure 9) and the stability of accuracy curves (Figure 8), the multi-scale attention integration is able to capture subtle but visually consistent patterns.

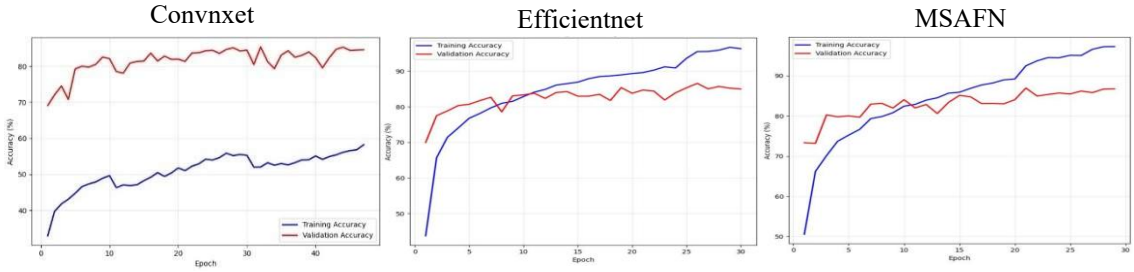


Figure 8: Training and validation accuracy curves for the proposed models on RAF-DB dataset

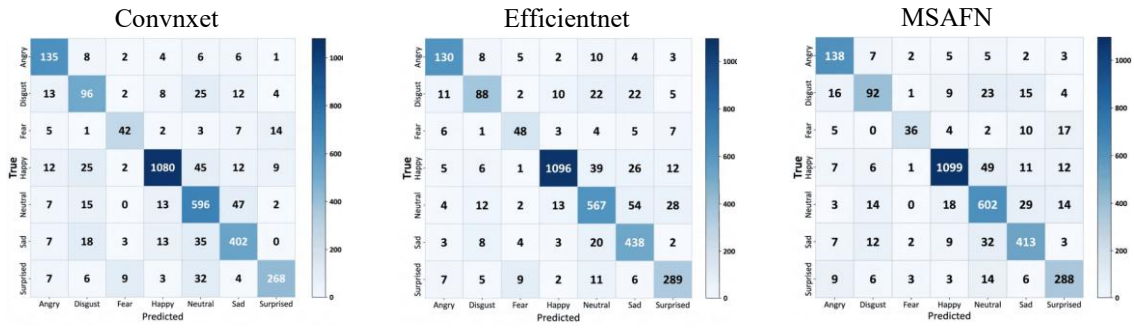


Figure 9: Confusion matrix for the proposed models on RAF-DB dataset

## *5.4. Comparative Evaluation of the Proposed Model and Recent State-of-the-Art Models*

In datasets with different levels of image quality and emotional complexity, the Multi-Scale Attention Fusion (MSAF-Fusion) model was always able to perform better at the task, especially when emotional expressions are operationalized by subtle and localized face muscle activity. ConvNeXt architecture showed consistent and reliable results on all the datasets which indicates that it has a high generalization capacity, but its sensitivity to subtle emotional variations was relatively less. In the meantime, EfficientNet worked well with enough detail of the face and clarity of the expression, but when the data was of low resolution, or the image was otherwise visually unclear, the performance decreased significantly.

To have a holistic assessment of the robustness of the proposed architecture, and demonstrate its excellence, a comparison analysis was made with the previous advanced studies as explained in the previous section. Such comparisons illustrate the clear benefits of the suggested Multi-Scale Attention Fusion (MSAFN) framework in comparison with traditional CNNs, hybrid CNN-Transformer systems, and attention-based systems. The outcomes of the suggested models were shown in Table 4, and these models were compared to the outcomes of related studies.As shown in Table 4, a comparison of the three benchmark datasets :FER-2013, RAF-DB, and AffectNet shows that the suggested models have significantly improved upon the other cutting-edge models that have been created to date.

When compared to spatial representation that could happen in low-resolution settings, the first convolutional networks proposed by Sarvakar et al. [16] using the FERC model scored only 54% in FER-2013. Designs that are lightweight, like the one by Gursesli et al. [22], achieved 65% performance with ShuffleNetV2,

and Mozaffari et al. [17] did 62% with a personalized CNN. Frameworks based on hybrids and transformers were then Singh, R. et al [36] introducing MoEffNet (EfficientNet-B0) with accuracy 81.68%.

Finally, despite the fact that these architectures improved spatial reasoning, they were limited by the coarse facial features and absence of context represented by FER-2013. Conversely MSAF model scored 72.18%. This regular enhancement suggests that multi-scale attention allows successful encoding of small muscle response and texture changes, even with low-resolution emotional pictures.

Table 4: Overall accuracy performance comparison for proposed models with different models

| **FER-2013** | **Accuracy** | **RAF-DB** | **Accuracy** | **AffectNet** | **Accuracy** |
|---|---|---|---|---|---|
| FERC[16] | 54% | DNN-CNN[34] | 65.67% | Effv2-s[35] | 57.80% |
| MoEffNet)[36] | 71.02% | ResNet-50[29] | 80% | HFE-Net[35] | 58.55% |
| HFE-Net[35] | 71.69% | DenseNet121[38] | 81% | POSTER-Var [37] | 67.91 % |
| Custom CNN[17] | 62% | MoEffNet 36] | 81.68% | Inception[19] | 58% |
| Inception-V3[19] | 65% | Custom CNN[18] | 84.32% | MobileNetV2[22] | 57% |
| ShuffleNetV2[22] | 65% | CLCM[22] | 84% | Custom CNN[18] | 61% |
| ConvNeXt (Proposed) | 71.06% | ConvNeXt (Proposed) | 85.37% | ConvNeXt (Proposed) | 77.58% |
| EfficientNet (Proposed) | 70.87% | EfficientNet (Proposed) | 86.57% | EfficientNet (Proposed) | 77.67% |
| MSAF (Proposed) | 72.18% | MSAF (Proposed) | 86.96% | MSAF(Proposed) | 78.93% |

The patterns of performance on RAF-DB exhibit the same trend but with a higher accuracy because it has richer texture and naturalistic variability. Huang et al. [34] documented a rate of 65.67% with a DNN-CNN method, whereas Chandra et al. [38] attained 81% by using DenseNet121, and Xie et al. [29] recorded 80% by applying ResNet-50 with attention mechanisms. More complex hybrid and transformer-based architectures, including Gomez-Sirvent et al. [18] obtained 84.32% with a custom CNN trained on in-the-wild data. However, the proposed model outperformed these models, with MSAFN achieving 86.96% accuracy. This result demonstrates the capability of the multi-scale attention to learn both the overall facial geometry and the local motion unit variations to achieve better generalization under unconstrained facial expression conditions. The challenge is much greater on the AffectNet dataset because of its high intra-class variability, a wide variety of poses, and environmental noise. These uncontrolled conditions were difficult to deal with in previous studies using CNNs and hybrid models. For example, Song and Liu [35] achieved an accuracy of 57.80% with the EffV2-S model and 58.55% with the HFE-Net architecture. Likewise, C.-T. Yen and K.-H. Li. The Inception model achieved a 58% accuracy on [19]. The difficulty of the dataset was demonstrated by the results of the POSTER-Var model by Gang et al. [37] which achieved 67.91 % and the MobileNetV2 by Gursesli et al. [22] with 57%. The proposed MSAFN model performed best, with 78.93% accuracy. The results validate the robustness of the proposed fusion strategy to the lighting variation, pose changes, and expression ambiguity by dynamically assigning weights to discriminative areas at different scales.This comparison as a whole justifies the excellence of the suggested MSAFN model. Although the previous studies conducted by Sarvakar et. al. [16], Huang et al. [28], and Chandra et al. [38] proposed incremental enhancements with enhanced architectures and attention mechanisms, multi-scale attention and feature fusion are a decisive move. MSAFN model was found to be more accurate than any other model on FER-2013, RAF-DB, and AffectNet, which validated its ability to resolve fine-grained, context-specific emotional cues that were distributed across heterogeneous facial areas.

## 6. Conclusion

A Multi-Scale Attention Fusion Network (MSAFN) with EfficientNet-B1 backbone was introduced in this study to achieve robust facial expression recognition. The main issue that our suggested solution will help solve is the need to cover both local fine-grained and the bigger contextual information with three parallel convolutional branches of 1x1, 3x3 and 5x5 kernel sizes. On FER2013, RAF-DB, and AffectNet, our MSAFN model was most accurate at 72.18%, 86.96%, and 78.93%. These results outperform premium models like ConvNeXt and EfficientNet-B1. Our results affirm that a combination of multi-scale attention and feature fusion can be used to offer greater discriminatory strength in order to identify subtle emotional expressions. The proposed architecture is successful in classifying subtle emotions like fear and surprise or sadness and disgust because it is able to capture both the global facial geometry and the localized action units. The work to come will involve incorporation of time in recognition of emotions based on videos and creating even simpler architectures that can

support real-time applications on mobile devices and at the edge, continuing to expand human-computer interaction systems.

## References


1. M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Front. Robot. AI*, vol. 7, p. 532279, 2020.
2. F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," *Soft Comput.*, vol. 28, no. 9–10, pp. 6695–6708, May 2024, doi:10.1007/s00500-023-09477-y.
3. Z. Shou, Y. Huang, D. Li, C. Feng, H. Zhang, Y. Lin, and G. Wu, "A student facial expression recognition model based on multi-scale and deep fine-grained feature attention enhancement," *Sensors*, vol. 24, no. 20, Art. no. 6748, 2024.
4. N. Lin and Y. Zuo, "Advancing driver fatigue detection in diverse lighting conditions for assisted driving vehicles with enhanced facial recognition technologies," *PLOS ONE*, vol. 19, no. 7, p. e0304669, July 2024, doi: 10.1371/journal.pone.0304669.
5. P. Ekman and W. V. Friesen, "Facial Action Coding System." 1978. doi: 10.1037/t27734-000.
6. S. Saeed, J. Baber, M. Bakhtyar, I. Ullah, N. Sheikh, I. Dad, and A. A. Sanjrani, "Empirical evaluation of SVM for facial expression recognition," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018.
7. B. K. Kumar, K. Swaroopa, and T. R. Balaga, "Facial emotion recognition and detection using CNN," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 14, pp. 5960–5968, 2021.
8. J. Daihong, Hu Yuanzheng, D. Lei, and P. Jin, "Facial Expression Recognition Based on Attention Mechanism," *Sci. Program.*, vol. 2021, pp. 1–10, Mar. 2021, doi: 10.1155/2021/6624251.
9. Y. Xi, C. Wu, T. Meng, and C. Li, "Facial emotion recognition based on ResNet18 with multi-dimensional attention mechanisms," *Memetic Comput.*, vol. 17, no. 4, p. 47, Dec. 2025, doi: 10.1007/s12293-025-00476-0.
10. K. Zu, H. Zhang, L. Zhang, J. Lu, C. Xu, H. Chen, and Y. Zheng, "EMBANet: A flexible efficient multi-branch attention network," *Neural Networks*, vol. 185, Art. no. 107248, 2025.
11. M. Koziarski, B. Kwolek, and B. Cyganek, "Convolutional neural network-based classification of histopathological images affected by data imbalance," in *Proceedings of the International Workshop on Face and Facial Expression Recognition from Real World Videos*, Cham, Switzerland: Springer, Aug. 2018, pp. 1–11.
12. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.
13. S. Li, W. Deng, and J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, July 2017, pp. 2584–2593. doi: 10.1109/CVPR.2017.277.
14. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, Berlin, Germany: Springer, Nov. 2013, pp. 117–124.
15. G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), 2018, pp. 423–430.
16. K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Mater. Today Proc.*, vol. 80, pp. 3560–3564, 2023, doi: 10.1016/j.matpr.2021.07.297.
17. L. Mozaffari, M. M. Brekke, B. Gajaruban, D. Purba, and J. Zhang, "Facial Expression Recognition Using Deep Neural Network," in *2023 3rd International Conference on Applied Artificial Intelligence (ICAPAI)*, Halden, Norway: IEEE, May 2023, pp. 1–9. doi: 10.1109/ICAPAI58366.2023.10193866.
18. J. L. Gómez-Sirvent, F. López De La Rosa, M. T. López, and A. Fernández-Caballero, "Facial Expression Recognition in the Wild for Low-Resolution Images Using Voting Residual Network," *Electronics*, vol. 12, no. 18, p. 3837, Sept. 2023, doi: 10.3390/electronics12183837.
19. C.-T. Yen and K.-H. Li, "Discussions of Different Deep Transfer Learning Models for Emotion Recognitions," *IEEE Access*, vol. 10, pp. 102860–102875, 2022, doi: 10.1109/ACCESS.2022.3209813.
20. S. Jyoti and A. Dhall, "Expression Empowered ResiDen Network for Facial Action Unit Detection," June 13, 2018, *arXiv*: arXiv:1806.04957. doi: 10.48550/arXiv.1806.04957.
21. A. A. Hameed Qutub and Y. Atay, "Deep Learning Approaches for Classification of Emotion Recognition based on Facial Expressions," *Nexo Rev. Científica*, vol. 36, no. 05, pp. 1–18, Nov. 2023, doi: 10.5377/nexo.v36i05.17181.
22. M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," *IEEE Access*, vol. 12, pp. 45543–45559, 2024, doi: 10.1109/ACCESS.2024.3380847.
23. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.
24. D. Amin, P. Chase, and K. Sinha, ''Touchy feely: An emotion recognition challenge,'' Palo Alto, Stanford, CA, USA, Tech. Rep., 2017.
25. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

26. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.
27. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.
28. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, July 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.
29. S. Xie, M. Li, S. Liu, and X. Tang, “*ResNet with Attention Mechanism and Deformable Convolution for Facial Expression Recognition*,” in 2021 4th International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China: IEEE, Sept. 2021, pp. 389–393. doi: 10.1109/ICICSP54369.2021.9611962.
30. M. Aminbeidokhti, M. Pedersoli, P. Cardinal, and E. Granger, “*Emotion Recognition with Spatial Attention and Temporal Softmax Pooling*,” vol. 11662, 2019, pp. 323–331. doi: 10.1007/978-3-030-27202-9_29.
31. J. Chen, L. Yang, L. Tan, and R. Xu, “Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition,” *Pattern Recognit.*, vol. 129, p. 108753, Sept. 2022, doi: 10.1016/j.patcog.2022.108753.
32. C. Ghadai, D. Patra, and M. Okade, “A novel facial expression recognition model based on harnessing complementary features in multi-scale network with attention fusion,” *Image Vis. Comput.*, vol. 149, p. 105183, Sept. 2024, doi: 10.1016/j.imavis.2024.105183.
33. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” 2022, *arXiv*. doi: 10.48550/ARXIV.2201.03545.
34. Z. Y. Huang, C. C. Chiang, J. H. Chen, Y. C. Chen, H. L. Chung, Y. P. Cai, and H. C. Hsu, “A study on computer vision for facial emotion recognition,” Scientific Reports, vol. 13, no. 1, Art. no. 8425, 2023.
35. D. Song and C. Liu, “A facial expression recognition network using hybrid feature extraction,” *PLOS ONE*, vol. 20, no. 1, p. e0312359, Jan. 2025, doi: 10.1371/journal.pone.0312359.
36. Singh, R. et al. (2022). EFFICIENTNET for human FER using transfer learning. ICTACT Journal on Soft Computing, 13(1).
37. Lv, G., Zhang, J., & Tsoi, C. (2026). Facial expression recognition via variational inference. *Scientific Reports*.
38. J. Chandra, A. B., and R. A. R., “Cross-Database Facial Expression Recognition using CNN with Attention Mechanism,” in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India: IEEE, July 2023, pp. 1–7. doi: 10.1109/ICCCNT56998.2023.10308238.